The Open University

M248

Analysing data

Handbook

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from Student Recruitment, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)300 303 5303; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

The Open University, Walton Hall, Milton Keynes, MK7 6AA.

First published 2017.

Edited, designed and typeset by The Open University, using the Open University TEX System.

WEB 05835 9

1.1

Contents

Introduction

This Handbook is provided as a reference document for M248. It provides a concise summary of the material in the units, a few mathematical results, tables of discrete and continuous distributions and their properties, and statistical tables for use in inferential procedures.

1 Unit summaries

Unit 1 Exploring and interpreting data

1. When describing a dataset, the following terminology is used:
   - **observations** (or **cases**, or **sampling units**) refer to objects (people, countries, ...) on which characteristics are recorded
   - **variables** are the characteristics recorded, and the pattern of variation of a variable is its **distribution**
   - variables are **linked** if they are each recorded for the same observations
   - a variable is **continuous** if its values are numerical and all values in an interval are possible
   - a variable is **discrete** if its values are numerical but only particular values (typically, integers) are possible
   - a variable is **categorical** if its values indicate to which group an observation belongs
   - a categorical variable is **ordinal** if its values correspond to labels which have a natural ordering
   - a categorical variable is **nominal** if its values correspond to labels but the labels do not have a natural ordering.
2. Useful graphical representations of data include bar charts, histograms, boxplots and scatterplots:
   - **bar charts** are generally used with categorical data, or with numerical data that are discrete; side-by-side bar charts can be used to display more than one such variable
   - **histograms** are generally used with continuous data; histograms come in frequency and unit-area versions which differ only in their 'vertical' scaling; histograms need a reasonably large dataset and are sensitive to the choice of cutpoints
   - **boxplots** are also generally used with continuous data; a comparative boxplot allows more than one continuous variable to be displayed at the same time; boxplots cannot show how many modes a distribution has
   - **scatterplots** are used to investigate the relationship between two numerical variables (which are often continuous but may be discrete).

3. A **mode** in a histogram corresponds to a peak in the heights of the bars. The data are **unimodal** if there is just one mode, **bimodal** if there are two modes and **multimodal** if there are more than two modes.
4. Numerical data that are not **symmetric**, in the sense that a bar chart or histogram shows a clear lack of symmetry, are said to be **skew**. If a bar chart or histogram has a relatively large 'tail' of relatively high values, then the data are **right-skew**; a dataset with a relatively long tail of relatively low values is **left-skew**.
5. If the n values in a dataset are denoted $x_1, x_2, \ldots, x_n$, then the **sample mean** is
$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$
6. When the data are written in order of increasing size, the pth value is denoted $x_{(p)}$. In particular:
   - the **sample median** is $m = x_{\left(\frac{1}{2}(n+1)\right)}$
   - the **sample lower quartile** is $q_L = x_{\left(\frac{1}{4}(n+1)\right)}$
   - the **sample upper quartile** is $q_U = x_{\left(\frac{3}{4}(n+1)\right)}$
   - the **sample interquartile range** is $q_U - q_L$.
7. The **sample standard deviation** is
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$
The quantity s^2 is the **sample variance**.

Unit 2 Modelling variation

1. If an experiment is repeated many times, then the number of times that an event E occurs is its **sample frequency** and the proportion of times that E occurs is its **sample relative frequency**. The **probability** that an event E occurs, $P(E)$, is the proportion towards which the sample relative frequency of E tends as the number of times the experiment is repeated increases.
2. Basic **properties of probabilities** include:
   - for any event E, $0 \leq P(E) \leq 1$
   - if an event E is impossible, then $P(E) = 0$
   - if an event E is certain to happen, then $P(E) = 1$
   - $P(E \text{ does not occur}) = 1 - P(E \text{ occurs})$
   - for **independent** events $E_1, E_2, \ldots, E_r$,
$$P(E_1 \text{ and } E_2 \text{ and } \ldots \text{ and } E_r) = P(E_1) \times P(E_2) \times \cdots \times P(E_r).$$

3. A random variable which takes only integer values is **discrete**. A **continuous** random variable may take any value within a continuous range of values.
4. The distribution of a discrete random variable X is given by its **probability mass function (p.m.f.)**, p:

   $$p(x) = P(X = x).$$

   For all x in the range of X, $0 < p(x) \leq 1$. Also, $\sum p(x) = 1$, where the sum is taken over all x in the range of X.
5. The distribution of a continuous random variable X is given by its **probability density function (p.d.f.)**, f. For all x in the range of X, $f(x) \geq 0$. Also, $\int f(x)\,dx = 1$, where the integral is taken over all x in the range of X.
6. For continuous X,

   $$P(x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} f(x)\,dx.$$

7. The **cumulative distribution function (c.d.f.)** of a random variable X is

   $$F(x) = P(X \leq x).$$

8. For continuous X with range $L < x < U$,

   $$F(x) = \int_{L}^{x} f(y)\,dy$$

   so that

   $$P(x_1 \leq X \leq x_2) = F(x_2) - F(x_1).$$

Unit 3 Models for discrete data

1. A **Bernoulli trial** is a single statistical experiment for which there are two possible outcomes, often referred to as success and failure and denoted by 1 and 0.
2. A random variable X with range $\{0, 1\}$ has a **Bernoulli distribution** with parameter p, where $0 < p < 1$, if it has p.m.f.

   $$p(0) = 1 - p, \quad p(1) = p.$$

   This is written $X \sim \text{Bernoulli}(p)$ and is the distribution of the outcome of a Bernoulli trial.
3. Two events are **independent** if the probability of the occurrence of one event is unaffected by whether or not the other occurs.

4. A random variable X has a **binomial distribution** with parameters n and p, where $0 < p < 1$, if it has p.m.f.

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n.$$

This is written $X \sim B(n, p)$. The binomial distribution provides a model for the total number of successes in a sequence of n independent Bernoulli trials, in which the probability of success in a single trial is p.

5. A random variable X has a **geometric distribution** with parameter p, where $0 < p < 1$, if it has p.m.f.

$$p(x) = (1-p)^{x-1} p, \quad x = 1, 2, 3, \ldots .$$

This is written $X \sim G(p)$. The geometric distribution provides a model for the number of trials up to and including the first success in a sequence of independent Bernoulli trials, in which the probability of success in each trial is p. The c.d.f. of X is

$$F(x) = 1 - (1-p)^x, \quad x = 1, 2, 3, \ldots .$$

6. A random variable X has a **Poisson distribution** with parameter λ, where $\lambda > 0$, if it has p.m.f.

$$p(x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \ldots .$$

This is written $X \sim \text{Poisson}(\lambda)$. The Poisson distribution is the limiting distribution of $X \sim B(n, \lambda/n)$ as n becomes large.

7. A random variable X has a **discrete uniform distribution** with parameters m and n, where $m < n$, if it has p.m.f.

$$p(x) = \frac{1}{n-m+1}, \quad x = m, m+1, \ldots, n.$$

This random variable is equally likely to take any integer value between m and n inclusive. The c.d.f. of X is

$$F(x) = \frac{x-m+1}{n-m+1}, \quad x = m, m+1, \ldots, n.$$

8. A random variable X has a **continuous uniform distribution** with parameters a and b, where $a < b$, if it has p.d.f.

$$f(x) = \frac{1}{b-a}, \quad a < x < b.$$

This is written $X \sim U(a, b)$. This random variable is equally likely to take any value between the two stated bounds. The c.d.f. of X is

$$F(x) = \frac{x-a}{b-a}, \quad a < x < b.$$

Unit 4 Population means and variances

1. The **population mean** (or **mean** or **expected value** or **expectation**) of a random variable is given:
   - if X is discrete with p.m.f. $p(x)$, by
     $$\mu = E(X) = \sum_x x\,p(x),$$
     where the sum is taken over all values x in the range of X
   - if X is continuous with p.d.f. $f(x)$, by
     $$\mu = E(X) = \int x\,f(x)\,dx,$$
     where the integral is taken over all values x in the range of X.
2. The **population variance** (or **variance**) of a random variable X is given:
   - if X is discrete with p.m.f. $p(x)$, by
     $$\sigma^2 = V(X) = E\big[(X-\mu)^2\big] = \sum_x (x-\mu)^2 p(x),$$
     where the sum is taken over all values x in the range of X
   - if X is continuous with p.d.f. $f(x)$, by
     $$\sigma^2 = V(X) = E\big[(X-\mu)^2\big] = \int (x-\mu)^2 f(x)\,dx,$$
     where the integral is taken over all values x in the range of X
   - whether X is discrete or continuous, also by
     $$\sigma^2 = V(X) = E(X^2) - \mu^2.$$
3. If X is a random variable and a and b are constants, then the mean and variance of $Y = aX + b$ are
   $$E(Y) = a\,E(X) + b, \quad V(Y) = a^2\,V(X).$$
4. If $X_1, X_2, \ldots, X_n$ are random variables, then
   $$E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$
   If $X_1, X_2, \ldots, X_n$ are independent random variables, then
   $$V(X_1 + X_2 + \cdots + X_n) = V(X_1) + V(X_2) + \cdots + V(X_n).$$
5. Formulas for the mean and variance of members of the families of distributions considered in this module are given in the tables of discrete and continuous probability distributions in this Handbook.

Unit 5 Events occurring at random and population quantiles

1. **Poisson's approximation for rare events** states that for large values of n and small values of p, the binomial random variable, $B(n, p)$, has approximately the same distribution as a Poisson random variable with parameter np:

   $$B(n, p) \approx \text{Poisson}(np).$$

   Equivalently, if $\mu = np$, then

   $$B(n, \mu/n) \approx \text{Poisson}(\mu).$$

   A rough rule for using Poisson's approximation is that:

   - if n is large and p is small ($n \geq 50$ and $p \leq 0.05$, say), then the approximation is good
   - when p is small enough, the approximation is good even for quite small values of n; the smallest value of n for which the approximation is good decreases as the value of p decreases.

2. A **Bernoulli process** is a sequence of Bernoulli trials in which:

   - trials are independent
   - the probability of success, p, remains the same from trial to trial.

   For a Bernoulli process:

   - the number of successes in n trials has a binomial distribution with parameters n and p
   - the waiting time from after one success up to and including the next success has a geometric distribution with parameter p.

3. A random variable X has an **exponential distribution** with parameter λ, where $\lambda > 0$, if it has p.d.f.

   $$f(x) = \lambda e^{-\lambda x}, \quad x > 0.$$

   This is written $X \sim M(\lambda)$. The c.d.f. of X is

   $$F(x) = 1 - e^{-\lambda x}, \quad x > 0.$$

4. The **Poisson process** is a model for the occurrence of events in continuous time in which:

   - events occur singly
   - the rate of occurrence of events remains constant
   - the incidence of future events is independent of the past.

   For a Poisson process in which events occur at random at rate λ:

   - the number of events that occur during a time interval of length t has a Poisson distribution with parameter λt
   - the waiting time between successive events has an exponential distribution with parameter λ.

5. The parallels between the Bernoulli and Poisson processes are summarised below.

Process	Type	Distribution of number of events	Distribution of waiting time between events
Bernoulli	discrete	binomial	geometric
Poisson	continuous	Poisson	exponential

6. For a continuous random variable X with c.d.f. $F(x)$, the **α-quantile** is the value x which is the solution of the equation

$$F(x) = \alpha, \quad 0 < \alpha < 1.$$

This value is denoted q_α.

7. For a discrete random variable X with c.d.f. $F(x)$, the **α-quantile**, q_α, is the smallest value of x in the range of X satisfying $F(x) \geq \alpha$.

8. The population **median**, m, **lower quartile**, q_L, and **upper quartile**, q_U, are those values of q_α corresponding to $\alpha = \frac{1}{2}$, $\frac{1}{4}$ and $\frac{3}{4}$, respectively.

Unit 6 Normal distributions

1. A random variable X has a **normal distribution** with mean μ and standard deviation σ (and hence variance σ^2), where $\sigma > 0$, if it has p.d.f.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}, \quad -\infty < x < \infty.$$

This is written $X \sim N(\mu, \sigma^2)$. The normal distribution is symmetric about μ.

2. If a normal distribution is used to model the variation in a population, then, according to the model, the proportion of the population within k standard deviations of the mean is the same, whatever the values of the mean μ and the standard deviation σ.

3. If $X \sim N(\mu, \sigma^2)$ and a and b are constants, then

$$Y = aX + b \sim N(a\mu + b, a^2\sigma^2).$$

4. If $X_1, X_2, \ldots, X_n$ are independent normally distributed random variables with means $\mu_1, \mu_2, \ldots, \mu_n$ and variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$, then

$$Y = X_1 + X_2 + \cdots + X_n \sim N(\mu_1 + \mu_2 + \cdots + \mu_n,\ \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2).$$

5. The normal distribution with mean 0 and standard deviation 1 is called the **standard normal distribution**. The letter Z is used to denote the standard normal random variable: $Z \sim N(0,1)$. The c.d.f. of Z is denoted $\Phi(z)$. The standard normal distribution is symmetric about 0. It follows that:
   - for any z,
     $$\Phi(-z) = 1 - \Phi(z)$$
   - if q_α is the α-quantile of Z, then for any $0 < \alpha < 1$,
     $$q_\alpha = -q_{1-\alpha}.$$
6. If $X \sim N(\mu, \sigma^2)$, then
   $$Z = \frac{X - \mu}{\sigma} \sim N(0,1).$$
   Conversely, if $Z \sim N(0,1)$, then
   $$X = \sigma Z + \mu \sim N(\mu, \sigma^2).$$
7. If $X \sim N(\mu, \sigma^2)$, then:
   - $P(X \leq x) = P\left(Z \leq \dfrac{x-\mu}{\sigma}\right) = \Phi\left(\dfrac{x-\mu}{\sigma}\right)$
   - the α-quantile, x, of X is
     $$x = \sigma q_\alpha + \mu,$$
     where q_α is the α-quantile of Z.

Table 1 of the statistical tables contains probabilities $\Phi(z)$ for the standard normal distribution.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

8. Given a set of n ordered observations $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, a **normal probability plot** is produced by plotting the n points $(x_{(i)}, y_i)$, $i = 1, 2, \ldots, n$, on a graph, where y_i is the quantile $q_{i/(n+1)}$ of a standard normal distribution. If the points lie roughly along a straight line, then a normal distribution is a plausible model for the variation in the data.
9. If X is a random variable with mean μ and variance σ^2, and if a random sample of size n is taken from the distribution of X, then the mean and variance of the sample total T_n are
   $$E(T_n) = n\mu, \quad V(T_n) = n\sigma^2,$$
   and the mean and variance of the sample mean $\overline{X}_n$ are
   $$E(\overline{X}_n) = \mu, \quad V(\overline{X}_n) = \frac{\sigma^2}{n}.$$
10. If $X_1, X_2, \ldots, X_n$ are n independent random observations from a population with mean μ and finite variance σ^2, then:
    - the **Central Limit Theorem** states that for large n, the distribution of their mean $\overline{X}_n$ is approximately normal with mean μ and variance σ^2/n:
      $$\overline{X}_n \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

- a corollary to the Central Limit Theorem states that for large n,

$$T_n = X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2).$$

11. For random samples of size n from a normal distribution with mean μ and variance σ^2, the sample total T_n and the sample mean $\overline{X}_n$ are exactly normally distributed:

$$T_n \sim N(n\mu, n\sigma^2), \quad \overline{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

Unit 7 Point estimation

1. There are many ways of obtaining **(point) estimators**, that is, estimating formulas, for unknown model parameters. When an estimating formula is applied in a data context, the resulting number provides a **(point) estimate** of the unknown parameter.
2. An estimator $\widehat{\theta}$ is said to be **unbiased** for a parameter θ if $E(\widehat{\theta}) = \theta$. An estimator $\widehat{\theta}$ is **biased** if it is not unbiased. The **bias** of $\widehat{\theta}$ is $E(\widehat{\theta}) - \theta$.
3. The sample mean $\overline{X}$ and sample variance S^2 are unbiased estimators of the population mean μ and the population variance σ^2, respectively.
4. If X is a discrete random variable with p.m.f. $p(x;\theta)$, then the **likelihood** for the random sample $x_1, x_2, \ldots, x_n$ is

$$L(\theta) = p(x_1;\theta) \times p(x_2;\theta) \times \cdots \times p(x_n;\theta).$$

   If X is a continuous random variable with p.d.f. $f(x;\theta)$, then the **likelihood** for the random sample $x_1, x_2, \ldots, x_n$ is

$$L(\theta) = f(x_1;\theta) \times f(x_2;\theta) \times \cdots \times f(x_n;\theta).$$

5. For a given sample of data, the **maximum likelihood estimate** $\widehat{\theta}$ of θ is the value of θ which maximises the likelihood. This estimate is an observation of the corresponding estimator, which is a random variable. This estimator is also denoted $\widehat{\theta}$ and is called the **maximum likelihood estimator** of θ. Both estimate and estimator are abbreviated to **MLE**.
6. Having formed the likelihood $L(\theta)$, a procedure for calculating the MLE $\widehat{\theta}$ of θ that is adequate for the problems discussed in this module is:
   - differentiate $L(\theta)$ to obtain $L'(\theta)$
   - solve the equation $L'(\theta) = 0$; if there is exactly one solution, then set $\widehat{\theta}$ equal to that solution.
7. MLEs possess the following properties:
   - they are sometimes unbiased and typically have small bias; also, they are **asymptotically unbiased**, that is,

$$E(\widehat{\theta}) \to \theta \text{ as } n \to \infty$$

- in addition,

  $$V(\widehat{\theta}) \to 0 \text{ as } n \to \infty.$$

  Moreover, for large n, no unbiased estimator of θ has a smaller variance than the MLE.

8. Formulas for the MLEs of the parameters of the families of distributions considered in this module are given in the tables of discrete and continuous probability distributions in this Handbook.

Unit 8 Interval estimation

1. Given a random sample $x_1, x_2, \ldots, x_n$ of size n from a population with mean μ, an approximate **$100(1-\alpha)\%$ confidence interval** for μ, valid for large n, is

   $$(\mu^-, \mu^+) = \left(\overline{x} - z\frac{s}{\sqrt{n}}, \overline{x} + z\frac{s}{\sqrt{n}}\right),$$

   where $\overline{x}$ is the sample mean, s is the sample standard deviation, and $z = q_{1-(\alpha/2)}$, the $(1-(\alpha/2))$-quantile of the standard normal distribution. The limits μ^- and μ^+ are, respectively, the **lower** and **upper $100(1-\alpha)\%$ confidence limits** for μ. This confidence interval is sometimes called a ***z*-interval**.

   Table 2 of the statistical tables contains quantiles for the standard normal distribution.

2. A **$100(1-\alpha)\%$ confidence interval** (θ^-, θ^+) for a population parameter θ, calculated from a sample of size n, may be interpreted as follows: if a large number of samples of size n were drawn independently from the population, and a $100(1-\alpha)\%$ confidence interval calculated on each occasion, then approximately $100(1-\alpha)\%$ of these intervals would contain the true parameter θ.

3. Suppose that (μ^-, μ^+) is a $100(1-\alpha)\%$ confidence interval for μ, and $\theta = h(\mu)$. If the transformation h is either increasing or decreasing, then the limits $h(\mu^-)$ and $h(\mu^+)$ define a $100(1-\alpha)\%$ confidence interval for θ as follows:

   - if h is increasing, then $(\theta^-, \theta^+) = (h(\mu^-), h(\mu^+))$
   - if h is decreasing, then $(\theta^-, \theta^+) = (h(\mu^+), h(\mu^-))$.

4. An **approximate $100(1-\alpha)\%$ confidence interval for a proportion** p, obtained by observing x successes in a sequence of n independent Bernoulli trials each with probability of success p, is

   $$(p^-, p^+) = \left(\widehat{p} - z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}, \widehat{p} + z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}\right),$$

   where $\widehat{p} = x/n$ and z is the $(1-(\alpha/2))$-quantile of the standard normal distribution. This confidence interval is valid when both np and $n(1-p)$ are at least 5.

5. Given observations x_1 and x_2 on independent binomial random variables $X_1 \sim B(n_1, p_1)$ and $X_2 \sim B(n_2, p_2)$, an **approximate $100(1-\alpha)\%$ confidence interval for the difference** $d = p_1 - p_2$ is

$$(d^-, d^+) = \left(\widehat{d} - z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}, \right.$$
$$\left. \widehat{d} + z\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}} \right),$$

where $\widehat{p}_1 = x_1/n_1$, $\widehat{p}_2 = x_2/n_2$, $\widehat{d} = \widehat{p}_1 - \widehat{p}_2$ and z is the $(1-(\alpha/2))$-quantile of the standard normal distribution.

6. An **approximate $100(1-\alpha)\%$ confidence interval for the Poisson parameter** λ is

$$(\lambda^-, \lambda^+) = \left(\overline{x} - z\sqrt{\frac{\overline{x}}{n}}, \overline{x} + z\sqrt{\frac{\overline{x}}{n}} \right),$$

where $\overline{x}$ is the sample mean, and z is the $(1-(\alpha/2))$-quantile of the standard normal distribution. This confidence interval is valid when $n\lambda$ is at least 30.

7. In a random sample of size n with sample mean $\overline{X}$ and sample standard deviation S from a normal distribution with mean μ, the random variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a ***t*-distribution** with $n-1$ degrees of freedom. This is written $T \sim t(n-1)$.

8. Given a random sample of size n with sample mean $\overline{x}$ and sample standard deviation s from a normal distribution with mean μ, a $100(1-\alpha)\%$ confidence interval for μ is

$$(\mu^-, \mu^+) = \left(\overline{x} - t\frac{s}{\sqrt{n}}, \overline{x} + t\frac{s}{\sqrt{n}} \right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-1)$. This confidence interval is exact and is sometimes called a ***t*-interval**.

Table 3 of the statistical tables contains quantiles for t-distributions.

9. Given independent samples of size n_1 with sample variance s_1^2 and n_2 with sample variance s_2^2 from distributions with a common variance, the pooled estimate of the common variance is

$$s_P^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}.$$

10. If n_1 and n_2 are the sample sizes, and $\overline{x}_1$ and $\overline{x}_2$ are the sample means of two independent samples from normal distributions with means μ_1 and μ_2 and common variance, then an exact $100(1-\alpha)\%$ confidence interval for the difference between the means, $d = \mu_1 - \mu_2$, is

$$(d^-, d^+) = \left(\overline{x}_1 - \overline{x}_2 - t\, s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},\ \overline{x}_1 - \overline{x}_2 + t\, s_P \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n_1+n_2-2)$ and s_P is the pooled estimate of the common standard deviation. This confidence interval is exact and is sometimes called a **two-sample *t*-interval**. The assumption of equal variances is valid if the larger of the two sample variances divided by the smaller is less than 3.

Unit 9 Testing hypotheses

1. The main steps in a 'fixed level' **hypothesis test** are:
   - set up the **null hypothesis**, H_0, and the **alternative hypothesis**, H_1
   - obtain some sample data and summarise these in the **test statistic**
   - obtain the **null distribution**; this is the distribution of the test statistic under the assumption that H_0 is true
   - decide on the **significance level** for the test; the significance level is the percentage of tests in which H_0 would be rejected when it is true and is usually one of 1%, 5% or 10%
   - calculate the **critical values** for the significance level, and hence the **rejection region** for the test; the latter, defined by the former, is the set of extreme values of the test statistic which lead to rejection of H_0
   - make one of two possible decisions:
     - **reject** H_0 if the test statistic lies in the rejection region
     - **do not reject** H_0 if the test statistic does not lie in the rejection region
   - state the conclusion of the test in non-technical language.
2. There are two commonly used tests for **testing a population mean**, μ, that is, testing

   $$H_0 : \mu = \mu_0$$

   against one of

   $$H_1 : \mu \neq \mu_0, \text{ or } H_1 : \mu < \mu_0, \text{ or } H_1 : \mu > \mu_0.$$

The ***z*-test**:

- can be used whatever the underlying distribution when the sample size is large ($n \geq 25$)
- the test statistic is

$$Z = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

- the null distribution is $N(0,1)$ so that critical values are found from the $N(0,1)$ quantile table.

Table 2 of the statistical tables contains quantiles for the standard normal distribution.

The ***t*-test**:

- can be used for a normal population for any sample size
- the test statistic is

$$T = \frac{\overline{X} - \mu_0}{S/\sqrt{n}}$$

- the null distribution is $t(n-1)$ so that critical values are found from the $t(n-1)$ quantile table.

Table 3 of the statistical tables contains quantiles for t-distributions.

3. To **test a population proportion** p, test

$$H_0 : p = p_0$$

against one of

$$H_1 : p \neq p_0, \text{ or } H_1 : p < p_0, \text{ or } H_1 : p > p_0.$$

The following test can be used when the sample size is large ($n \geq 25$):

- the test statistic is

$$Z_p = \frac{\frac{X}{n} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

- the null distribution is $N(0,1)$ so that critical values are found from the $N(0,1)$ quantile table.

4. The main steps in using p-values for testing hypotheses are:
   - set up the null and alternative hypotheses
   - obtain some sample data and summarise these in the test statistic
   - obtain the null distribution of the test statistic
   - identify all other values of the test statistic that are at least as extreme, in relation to the null and alternative hypotheses, as the value that was observed
   - using the null distribution, calculate the ***p*-value** as the probability of observing a value of the test statistic at least as extreme as the value observed
   - interpret the p-value
   - state the conclusion of the test in non-technical language.

5. The table below provides a rough guide to interpreting p-values.

p-value	Rough interpretation
$p > 0.10$	little or no evidence against H_0
$0.05 < p \leq 0.10$	weak evidence against H_0
$0.01 < p \leq 0.05$	moderate evidence against H_0
$p \leq 0.01$	strong evidence against H_0

6. Suppose that a hypothesis test results in a p-value p. Then a hypothesis test at significance level $100\alpha\%$ would result in:
   - the null hypothesis being rejected if $p \leq \alpha$
   - the null hypothesis not being rejected if $p > \alpha$.
7. The conclusions of a hypothesis test may be in error:
   - a **Type I error** occurs when we reject H_0 but it is true then
     $$\text{significance level} = P(\text{Type I error})$$
   - a **Type II error** occurs when we do not reject H_0 but it is false
   - there is a trade-off between the two error probabilities when designing a test.
8. The **power** of a test is
   $$\text{power} = P(\text{reject } H_0 \text{ when } H_0 \text{ is false}) = 1 - P(\text{Type II error}).$$
   It is desirable to have large power and small significance level.
9. Suppose that a sample of size n is obtained from a population distributed as $N(\mu, \sigma^2)$, where σ^2 is known, and the test statistic
   $$Z_1 = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}$$
   is to be used in a test of $H_0 : \mu = \mu_0$ with significance level α. Let $d > 0$.
   - When the alternative hypothesis is $H_1 : \mu > \mu_0$ and the true value of μ is $\mu_0 + d$ or when the alternative hypothesis is $H_1 : \mu < \mu_0$ and the true value of μ is $\mu_0 - d$, then the **power of the one-sided test** is
     $$1 - \Phi\left(q_{1-\alpha} - \frac{d}{\sigma/\sqrt{n}}\right).$$
   - When the alternative hypothesis is $H_1 : \mu \neq \mu_0$ and the true value of μ is $\mu_0 \pm d$, where d is not small, then the **power of the two-sided test** is approximately
     $$1 - \Phi\left(q_{1-(\alpha/2)} - \frac{d}{\sigma/\sqrt{n}}\right).$$

10. Suppose that a sample of data is to be collected and one of the tests described in the previous point is to be performed. Suppose also that n is to be chosen so that the power of the test, when the true underlying mean is $\mu_0 + d$, is equal to a predetermined value γ. The required sample size is:
    - for a one-sided test,
    $$n = \frac{\sigma^2}{d^2}\left(q_{1-\alpha} - q_{1-\gamma}\right)^2$$
    - for a two-sided test,
    $$n = \frac{\sigma^2}{d^2}(q_{1-(\alpha/2)} - q_{1-\gamma})^2.$$

Unit 10 Nonparametric and goodness-of-fit tests

1. The **Wilcoxon signed rank test** is a test on a single sample of data, $x_1, x_2, \ldots, x_{n_1}$. Let m denote the underlying population. To test
    $$H_0 : m = m_0$$
    against one of
    $$H_1 : m \neq m_0, \text{ or } H_1 : m < m_0, \text{ or } H_1 : m > m_0,$$
    the following test can be used:
    - – if the data are a set of paired differences, then we are testing for a zero median, so $m_0 = 0$; set $d_i = x_i$, $i = 1, 2, \ldots, n_1$
      – for a single sample for which we are testing for a non-zero median, $m_0 \neq 0$, form the differences from the specified value m_0: $d_i = x_i - m_0$, $i = 1, 2, \ldots, n_1$
    - in either case, delete any zeros from the dataset of differences and let $n \leq n_1$ be the sample size of the dataset with zeros removed
    - rewrite the null and alternative hypotheses with $m_0 = 0$
    - without regard to their sign, order the absolute values of the differences from least to greatest, and allocate rank i to the ith absolute difference; in the event of ties, allocate the average rank to the tied differences
    - consider again the signs of the original differences; the Wilcoxon signed rank test statistic, w_+, is the sum of the ranks of the positive differences
    - obtain the p-value, p
    - interpret p and state your conclusions.
2. Under the null hypothesis, for a sample of size n (excluding any zero differences), the Wilcoxon signed rank test statistic W_+ has mean and variance given by
    $$E(W_+) = \frac{n(n+1)}{4}, \quad V(W_+) = \frac{n(n+1)(2n+1)}{24}.$$

The distribution of

$$Z = \frac{W_+ - E(W_+)}{\sqrt{V(W_+)}}$$

is approximately standard normal. The approximation is adequate provided that $n \geq 16$.

3. The **Mann–Whitney test** is a test on two independent samples of data. Let ℓ denote the underlying difference in location between the populations from which the samples were drawn. To test

   $$H_0 : \ell = 0$$

   against one of

   $$H_1 : \ell \neq 0, \text{ or } H_1 : \ell < 0, \text{ or } H_1 : \ell > 0,$$

   the following test can be used:

   - pool the two samples, keeping track of the sample to which each data value belongs
   - order the pooled data values from least to greatest, and allocate rank i to the ith pooled value; in the event of ties, allocate the average rank to the tied values
   - the Mann–Whitney test statistic, u_A, is the sum of the ranks for one of the samples
   - obtain the p-value, p
   - interpret p and state your conclusions.

4. For independent samples of sizes n_A and n_B, the null distribution of the Mann–Whitney test statistic U_A has mean and variance given by

   $$E(U_A) = \frac{n_A\,(n_A + n_B + 1)}{2}, \quad V(U_A) = \frac{n_A n_B\,(n_A + n_B + 1)}{12}.$$

   The distribution of

   $$Z = \frac{U_A - E(U_A)}{\sqrt{V(U_A)}}$$

   is approximately standard normal. The approximation can be used if $n_A \geq 8$ and $n_B \geq 8$ and there are not too many tied values in the pooled dataset.

5. The random variable W given by the sum of the squares of r independent standard normal random variables has a **chi-squared distribution** with r degrees of freedom. This is written $W \sim \chi^2(r)$.

6. Given a random sample for which each observation can be classified into one of k distinct categories, the **chi-squared goodness-of-fit test** involves the comparison of the observed frequencies for the categories and the frequencies expected under a hypothesised model. The test statistic is

   $$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i},$$

where O_i and E_i are the observed and expected frequencies for category i. The categories must be chosen in such a way that the expected frequency for each category is at least 5. Then, under the null hypothesis that the data arise from the hypothesised model, the distribution of the test statistic χ^2 is approximately chi-squared with $k-p-1$ degrees of freedom, where p is the number of parameters whose values were estimated from the data. The p-value is given by the upper tail probability of $\chi^2(k-p-1)$ for values exceeding the observed test statistic.

Table 4 of the statistical tables contains quantiles for chi-squared distributions.

Unit 11 Regression

1. When a **response variable** Y is related to the value of an **explanatory variable** x, then the relationship can be represented by a **general regression model**

$$Y_i = h(x_i) + W_i, \quad i = 1, 2, \ldots, n.$$

Here h represents some **regression function** and the W_is are independent random variables with zero mean.

2. An important regression model is the **(simple) linear regression model**, where Y depends linearly on x, that is,

$$Y_i = \alpha + \beta x_i + W_i, \quad i = 1, 2, \ldots, n.$$

The line $y = \alpha + \beta x$ is called the **regression line**, with parameters α being the **intercept** and β the **slope**. The random terms W_i are independent with zero mean and constant variance σ^2. Often, the W_i are additionally assumed to be normally distributed.

3. The following notation is standard:

$$S_{xx} = \sum (x_i - \overline{x})^2 = \sum x_i^2 - \frac{\left(\sum x_i\right)^2}{n},$$
$$S_{yy} = \sum (y_i - \overline{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n},$$
$$S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}) = \sum x_i y_i - \frac{\left(\sum x_i\right)\left(\sum y_i\right)}{n}.$$

4. Given data, the parameters of the linear regression model may be estimated using the **method of least squares** by minimising the sum of squared residuals

$$\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2.$$

- The **least squares estimate** of β is

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}}.$$

- The least squares estimate of α is

$$\widehat{\alpha} = \overline{y} - \widehat{\beta}\overline{x}.$$

- The **least squares line** is

$$y = \widehat{\alpha} + \widehat{\beta}\, x.$$

5. The assumption that the W_i have constant, zero mean and constant variance can be checked using a **residual plot** in which the **residuals** $w_i = y_i - \widehat{y}_i$ are plotted against the **fitted values** $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta} x_i$. If this assumption is satisfied, then the further assumption of normality of the W_i can be checked using a **normal probability plot of the residuals**.

6. An unbiased estimator of σ^2 is

$$S^2 = \frac{\sum (Y_i - \widehat{Y}_i)^2}{n-2}.$$

7. Assuming that, independently, $W_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$, we have

$$\widehat{\beta} \sim N\left(\beta, \frac{\sigma^2}{S_{xx}}\right), \quad \frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2).$$

These two results can be combined to give

$$\frac{\widehat{\beta} - \beta}{S/\sqrt{S_{xx}}} \sim t(n-2).$$

Table 3 of the statistical tables contains quantiles for t-distributions.

This result can be used to test $H_0 : \beta = 0$.

8. If the random terms W_i are normally distributed, then a **$100(1-\alpha)\%$ confidence interval for the parameter β** is

$$\left(\widehat{\beta} - t\frac{s}{\sqrt{S_{xx}}}, \widehat{\beta} + t\frac{s}{\sqrt{S_{xx}}}\right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

9. If the random terms W_i are normally distributed, then, for a given x_0, a **$100(1-\alpha)\%$ confidence interval for the mean response** is

$$\left(\widehat{\alpha} + \widehat{\beta}x_0 - t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}},\ \widehat{\alpha} + \widehat{\beta}x_0 + t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n}}\right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

10. If the random terms W_i are normally distributed, then, for a given x_0, a **$100(1-\alpha)\%$ prediction interval for the response** is

$$\left(\widehat{\alpha} + \widehat{\beta}x_0 - t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n} + 1},\ \widehat{\alpha} + \widehat{\beta}x_0 + t\,s\sqrt{\frac{(x_0-\overline{x})^2}{S_{xx}} + \frac{1}{n} + 1}\right),$$

where t is the $(1-(\alpha/2))$-quantile of $t(n-2)$.

11. If data comprise observations on p explanatory variables $x_1, x_2, \ldots, x_p$ and a response variable Y, then the **multiple linear regression model** can be written

$$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + W_i, \quad i = 1, 2, \ldots, n.$$

The terms W_i are independent normal random variables with zero mean and constant variance.

Unit 12 Transformations and the modelling process

1. The **ladder of powers** lists transformations of the form

   $$\ldots,\ x^{-2},\ x^{-1},\ x^{-1/2},\ \log x,\ x^{1/2},\ x^{1},\ x^{2},\ x^{3},\ x^{4},\ \ldots\ .$$

   When transforming skew, positive data to make them more symmetric, and hence more amenable to modelling with a normal distribution:

   - for right-skew data, go down the ladder of powers
   - for left-skew data, go up the ladder of powers.

2. In linear regression, it is sometimes possible to:

   - straighten out the regression function by **transforming the explanatory variable**
   - make the assumptions associated with the random terms conform to those of the linear regression model by **transforming the response variable**.

3. A **statistical report** comprises the following sections: *Summary*, *Introduction*, *Methods*, *Results*, *Discussion*:

   - the *Summary* should be self-contained and should be written in largely non-technical language; it should state briefly the aim of the analysis, the methods used, the key finding(s), and the interpretation
   - the *Introduction* should contain a brief description of the question or hypothesis to be investigated, the setting in which the data were collected, and the data available
   - the *Methods* section should include a description of the model, the procedures used to check the model, the statistical tests employed, the methods used to calculate confidence intervals, and any other relevant techniques used
   - the *Results* section should contain descriptive summaries of the data, evidence that the model is appropriate, and the numerical results of statistical tests and confidence interval calculations
   - the *Discussion* should contain your assessment of the statistical evidence relating to the original question or hypothesis.

Unit 13 Applications

This unit uses techniques from the previous units to solve applied problems.

2 Some mathematical results

1. $\binom{n}{x} = \dfrac{n!}{x!\,(n-x)!}$, where $x! = 1 \times 2 \times \cdots \times x$ and $0! = 1$.
2. Integration rules include:
   - for $k \neq -1$,
     $$\int ax^k \, dx = \frac{ax^{k+1}}{(k+1)} + c$$
   - if $g(x), h(x), \ldots, q(x)$ are any functions of x, then
     $$\int \{g(x) + h(x) + \cdots + q(x)\} \, dx$$
     $$= \int g(x) \, dx + \int h(x) \, dx + \cdots + \int q(x) \, dx$$
   - if $g(x)$ is any function of x and a is a constant, then
     $$\int a\, g(x) \, dx = a \int g(x) \, dx$$
   - if $f(x)$ is a function such that $\int f(x) \, dx = G(x) + c$, and $x_1 < x_2$, then
     $$\int_{x_1}^{x_2} f(x) \, dx = \Big[G(x)\Big]_{x_1}^{x_2} = G(x_2) - G(x_1)$$
   - for $a \neq 0$,
     $$\int e^{ax} \, dx = \frac{e^{ax}}{a} + c.$$
3. Differentiation rules include:
   - if $f(x) = ax^k$, then $\frac{d}{dx} f(x) = f'(x) = kax^{k-1}$
   - let $g(x), h(x), \ldots, q(x)$ be any functions of x, and $a, b, \ldots, k$ be constants:
     $$\text{if } f(x) = a\, g(x) + b\, h(x) + \cdots + k\, q(x),$$
     $$\text{then } f'(x) = a\, g'(x) + b\, h'(x) + \cdots + k\, q'(x)$$
   - if $f(x) = ae^{kx}$, then $f'(x) = kae^{kx}$
   - the chain rule:
     $$\text{if } f(x) = h(g(x)) \text{ for suitable functions } g \text{ and } h,$$
     $$\text{then } f'(x) = g'(x)\, h'(g(x))$$
   - the product rule:
     $$\text{if } f(x) = g(x) \times h(x) \text{ for any functions } g \text{ and } h,$$
     $$\text{then } f'(x) = g'(x)\, h(x) + g(x)\, h'(x).$$

3 Table of discrete probability distributions

Name and abbreviation	Probability mass function $p(x)$	Cumulative distribution function $F(X)$	Range	Parameter values	Mean μ	Variance σ^2	Maximum likelihood estimator of parameter
Bernoulli Bernoulli(p)	$p(0) = 1 - p,$ $p(1) = p$		$0, 1$	$0 < p < 1$	p	$p(1-p)$	
Binomial $B(n, p)$	$\binom{n}{x} p^x (1-p)^{n-x}$		$0, 1, \ldots, n$	$n = 1, 2, 3, \ldots,$ $0 < p < 1$	np	$np(1-p)$	$\widehat{p} = \frac{X}{n}$
Discrete uniform	$\frac{1}{n-m+1}$	$\frac{x-m+1}{n-m+1}$	$m, m+1, \ldots, n$	$m < n$ both whole numbers	$\frac{n+m}{2}$	$\frac{(n-m)(n-m+2)}{12}$	
Geometric $G(p)$	$(1-p)^{x-1} p$	$1 - (1-p)^x$	$1, 2, 3, \ldots$	$0 < p < 1$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\widehat{p} = \frac{1}{\overline{X}}$
Poisson Poisson(λ)	$\frac{e^{-\lambda} \lambda^x}{x!}$		$0, 1, 2, \ldots$	$\lambda > 0$	λ	λ	$\widehat{\lambda} = \overline{X}$

4 Table of continuous probability distributions

Name and abbreviation	Probability density function $f(x)$	Cumulative distribution function $F(X)$	Range	Parameter values	Mean μ	Variance σ^2	Maximum likelihood estimator(s) of parameter(s)
Chi-squared $\chi^2(r)$			$x > 0$	$r = 1, 2, 3, \ldots$	r	$2r$	
Continuous uniform $U(a, b)$	$\frac{1}{b-a}$	$\frac{x-a}{b-a}$	$a < x < b$	$-\infty < a < b < \infty$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	
Exponential $M(\lambda)$	$\lambda e^{-\lambda x}$	$1 - e^{-\lambda x}$	$x > 0$	$\lambda > 0$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\widehat{\lambda} = \frac{1}{\overline{X}}$
Normal $N(\mu, \sigma^2)$	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\}$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$	$-\infty < x < \infty$	$-\infty < \mu < \infty$, $\sigma > 0$	μ	σ^2	$\widehat{\mu} = \overline{X}$, $\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}$
Standard normal $N(0, 1)$	$\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$	$\Phi(x)$	$-\infty < x < \infty$		0	1	
Student's t $t(\nu)$	proportional to $\left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$		$-\infty < x < \infty$	$\nu = 1, 2, 3, \ldots$			

5 Statistical tables

Table 1 Probabilities for the standard normal distribution $\Phi(z) = P(Z \leq z)$

z	0	1	2	3	4	5	6	7	8	9
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
4.0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Example: $\Phi(1.58) = 0.9429$.

Table 2 Quantiles for the standard normal distribution, $\Phi(q_\alpha) = \alpha$

α	q_α	α	q_α	α	q_α	α	q_α
0.50	0.00000	0.67	0.4399	0.84	0.9945	0.955	1.695
0.51	0.02507	0.68	0.4677	0.85	1.036	0.960	1.751
0.52	0.05015	0.69	0.4959	0.86	1.080	0.965	1.812
0.53	0.07527	0.70	0.5244	0.87	1.126	0.970	1.881
0.54	0.1004	0.71	0.5534	0.88	1.175	0.975	1.960
0.55	0.1257	0.72	0.5828	0.89	1.227	0.980	2.054
0.56	0.1510	0.73	0.6128	0.90	1.282	0.985	2.170
0.57	0.1764	0.74	0.6433	0.905	1.311	0.990	2.326
0.58	0.2019	0.75	0.6745	0.910	1.341	0.991	2.366
0.59	0.2275	0.76	0.7063	0.915	1.372	0.992	2.409
0.60	0.2533	0.77	0.7388	0.920	1.405	0.993	2.457
0.61	0.2793	0.78	0.7722	0.925	1.440	0.994	2.512
0.62	0.3055	0.79	0.8064	0.930	1.476	0.995	2.576
0.63	0.3319	0.80	0.8416	0.935	1.514	0.996	2.652
0.64	0.3585	0.81	0.8779	0.940	1.555	0.997	2.748
0.65	0.3853	0.82	0.9154	0.945	1.598	0.998	2.878
0.66	0.4125	0.83	0.9542	0.950	1.645	0.999	3.090

Example: $q_{0.950} = 1.645$.

Table 3 Quantiles for t-distributions

df	0.90	0.95	0.975	0.99	0.995	0.999
1	3.078	6.314	12.71	31.82	63.66	318.3
2	1.886	2.920	4.303	6.965	9.925	22.33
3	1.638	2.353	3.182	4.541	5.841	10.21
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
31	1.309	1.696	2.040	2.453	2.744	3.375
32	1.309	1.694	2.037	2.449	2.738	3.365
33	1.308	1.692	2.035	2.445	2.733	3.356
34	1.307	1.691	2.032	2.441	2.728	3.348
35	1.306	1.690	2.030	2.438	2.724	3.340
36	1.306	1.688	2.028	2.434	2.719	3.333
37	1.305	1.687	2.026	2.431	2.715	3.326
38	1.304	1.686	2.024	2.429	2.712	3.319
39	1.304	1.685	2.023	2.426	2.708	3.313
40	1.303	1.684	2.021	2.423	2.704	3.307
45	1.301	1.679	2.014	2.412	2.690	3.281
50	1.299	1.676	2.009	2.403	2.678	3.261
55	1.297	1.673	2.004	2.396	2.668	3.245
60	1.296	1.671	2.000	2.390	2.660	3.232
65	1.295	1.669	1.997	2.385	2.654	3.220
70	1.294	1.667	1.994	2.381	2.648	3.211
75	1.293	1.665	1.992	2.377	2.643	3.202
80	1.292	1.664	1.990	2.374	2.639	3.195
85	1.292	1.663	1.988	2.371	2.635	3.189
90	1.291	1.662	1.987	2.368	2.632	3.183
100	1.290	1.660	1.984	2.364	2.626	3.174

Example: $P(T \leq 2.262) = 0.975$, where $T \sim t(9)$.

Table 4 Quantiles for χ^2 distributions

df	0.005	0.01	0.025	0.05	0.1	0.5	0.9	0.95	0.975	0.99	0.995
1	0.00004	0.0001	0.0009	0.0039	0.016	0.455	2.71	3.84	5.02	6.63	7.88
2	0.010	0.020	0.051	0.103	0.211	1.39	4.61	5.99	7.38	9.21	10.60
3	0.072	0.115	0.216	0.352	0.584	2.37	6.25	7.81	9.35	11.34	12.84
4	0.207	0.297	0.484	0.711	1.06	3.36	7.78	9.49	11.14	13.28	14.86
5	0.412	0.554	0.831	1.14	1.61	4.35	9.24	11.07	12.83	15.09	16.75
6	0.676	0.872	1.24	1.64	2.20	5.35	10.64	12.59	14.45	16.81	18.55
7	0.989	1.24	1.69	2.17	2.83	6.35	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	7.34	13.36	15.51	17.53	20.09	21.95
9	1.73	2.09	2.70	3.33	4.17	8.34	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	9.34	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	10.34	17.28	19.68	21.92	24.72	26.76
12	3.07	3.57	4.40	5.23	6.30	11.34	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	12.34	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	13.34	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	14.34	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	15.34	23.54	26.30	28.85	32.00	34.27
17	5.70	6.41	7.56	8.67	10.09	16.34	24.77	27.59	30.19	33.41	35.72
18	6.26	7.01	8.23	9.39	10.86	17.34	25.99	28.87	31.53	34.81	37.16
19	6.84	7.63	8.91	10.12	11.65	18.34	27.20	30.14	32.85	36.19	38.58
20	7.43	8.26	9.59	10.85	12.44	19.34	28.41	31.41	34.17	37.57	40.00
21	8.03	8.90	10.28	11.59	13.24	20.34	29.62	32.67	35.48	38.93	41.40
22	8.64	9.54	10.98	12.34	14.04	21.34	30.81	33.92	36.78	40.29	42.80
23	9.26	10.20	11.69	13.09	14.85	22.34	32.01	35.17	38.08	41.64	44.18
24	9.89	10.86	12.40	13.85	15.66	23.34	33.20	36.42	39.36	42.98	45.56
25	10.52	11.52	13.12	14.61	16.47	24.34	34.38	37.65	40.65	44.31	46.93
26	11.16	12.20	13.84	15.38	17.29	25.34	35.56	38.89	41.92	45.64	48.29
27	11.81	12.88	14.57	16.15	18.11	26.34	36.74	40.11	43.19	46.96	49.64
28	12.46	13.56	15.31	16.93	18.94	27.34	37.92	41.34	44.46	48.28	50.99
29	13.12	14.26	16.05	17.71	19.77	28.34	39.09	42.56	45.72	49.59	52.34
30	13.79	14.95	16.79	18.49	20.60	29.34	40.26	43.77	46.98	50.89	53.67
31	14.46	15.66	17.54	19.28	21.43	30.34	41.42	44.99	48.23	52.19	55.00
32	15.13	16.36	18.29	20.07	22.27	31.34	42.58	46.19	49.48	53.49	56.33
33	15.82	17.07	19.05	20.87	23.11	32.34	43.75	47.40	50.73	54.78	57.65
34	16.50	17.70	19.81	21.66	23.95	33.34	44.90	48.60	51.97	56.06	58.96
35	17.19	18.51	20.57	22.47	24.80	34.34	46.06	49.80	53.20	57.34	60.27
36	17.89	19.23	21.34	23.27	25.64	35.34	47.21	51.00	54.44	58.62	61.58
37	18.59	19.96	22.11	24.07	26.49	36.34	48.36	52.19	55.67	59.89	62.88
38	19.29	20.69	22.88	24.88	27.34	37.34	49.51	53.38	56.90	61.16	64.18
39	20.00	21.43	23.65	25.70	28.20	38.34	50.66	54.57	58.12	62.43	65.48
40	20.71	22.16	24.43	26.51	29.05	39.34	51.81	55.76	59.34	63.69	66.77
45	24.31	25.90	28.37	30.61	33.35	44.34	57.51	61.66	65.41	69.96	73.17
50	27.99	29.71	32.36	34.76	37.69	49.33	63.17	67.50	71.42	76.15	79.49
55	31.73	33.57	36.40	38.96	42.06	54.33	68.80	73.31	77.38	82.29	85.75
60	35.53	37.48	40.48	43.19	46.46	59.33	74.40	79.08	83.30	88.38	91.95
65	39.38	41.44	44.60	47.45	50.88	64.33	79.97	84.82	89.18	94.42	98.11
70	43.28	45.44	48.76	51.74	55.33	69.33	85.53	90.53	95.02	100.4	104.2
75	47.21	49.48	52.94	56.05	59.79	74.33	91.06	96.22	100.8	106.4	110.3
80	51.17	53.54	57.15	60.39	64.28	79.33	96.58	101.9	106.6	112.3	116.3
85	55.17	57.63	61.39	64.75	68.78	84.33	102.1	107.5	112.4	118.2	122.3
90	59.20	61.75	65.65	69.13	73.29	89.33	107.6	113.1	118.1	124.1	128.3
100	67.33	70.06	74.22	77.93	82.36	99.33	118.5	124.3	129.6	135.8	140.2

Example: $P(W \leq 33.92) = 0.95$, where $W \sim \chi^2(22)$.